Preprint typeset in JHEP style - HYPER VERSION

Pure Samples of Quark and Gluon Jets at the LHC

Jason Gallicchio and Matthew D. Schwartz

Jefferson Physical Laboratory, Harvard University, Cambridge, MA 02138

Abstract: Having pure samples of quark and gluon jets would greatly facilitate the study of jet properties and substructure, with many potential standard model and new physics applications. To this end, we consider multijet and jets+X samples, to determine the purity that can be achieved by simple kinematic cuts leaving reasonable production cross sections. We find, for example, that at the 7 TeV LHC, the $pp \to \gamma$+2jets sample can provide 98% pure quark jets with 200 GeV of transverse momentum and a cross section of 5 pb. To get 10 pb of 200 GeV jets with 90% gluon purity, the $pp \to$ 3jets sample can be used. b+2jets is also useful for gluons, but only if the b-tagging is very efficient.

1. Introduction

Proton colliders, like the Large Hadron Collider at CERN, produce an enormous number of high energy jets. These jets are manifestations of hard quarks or gluons produced at very short distances, which shower and fragment into collections of collinear particles. Being able to distinguish quark and gluon jets could be extremely useful for new physics searches. For example, many models with supersymmetry produce dominantly quark jets while their backgrounds are dominantly gluon jets. The hope is then to discriminate signal from background by using observables like jet mass, which are strongly correlated with flavor [1, 2, 3, 4, 5, 6, 7, 8]. In order to validate these observables on data, it would be useful to have relatively pure samples of light quark or gluon jets to study. It is the purpose of this paper to suggest where those samples might be found.

At leading order in perturbation theory, there is no ambiguity in what is meant by the quark and gluon jet fraction in any exclusive sample. For example, as we show below, in a 300 GeV dijet sample at the 7 TeV LHC, the division is roughly 50/50. This comes simply from the ratio of the LO cross sections for the various channels, which do not interfere. The fraction can be defined beyond leading-order as well. In fact, it is well-defined to to all orders in perturbation theory up to the same power corrections that affect any jet algorithm's parton correspondence. These power corrections involve the jet size R (equivalently the jet's mass-to-energy ratio m/E) and $\Lambda_{\rm QCD}/E$. One can also define an infrared-safe definition of flavor at the jet level [9], but that is not the subject of this paper. We further discuss the theoretical issues associated with defining quark and gluon jets in Section 4.

To be clear, we do not propose that the quark and gluon fractions can be measured directly in data. Instead, one can measure observable properties of the samples, such as the jet mass, and compare them to theoretical predictions, such as from Monte Carlo simulations. The purity calculations in this paper suggest regions where the measurements would be most enlightening.

It may not be obvious why one would want pure samples of quark or gluon jets at all. Instead, one could just study the jet observables directly in any mixed sample. For example, it is well known that the distribution of jet mass for 300 GeV jets is typically wider and peaks at larger values for gluon jets than quark jets. In a 50/50 sample, such as the 300 GeV dijet sample, one could then to hope to find two separated peaks. Unfortunately, the combined distribution does *not* have two distinct peaks for jet mass, or charged particle count, or any other known discriminant — the distributions are just too broad. Moreover, correlations in the 2D distribution of observables like jet mass and charged particle count might take different forms that would be impossible to see in a 50/50 sample. The purer the sample, the closer one can come to studying quark and gluon jets on an event-by-event basis.

In this paper, we simulate a wide variety of processes at tree level for the 7 TeV LHC. These include events with gluon and light quark (uds) jets, b-jets, W's, Z's and γ's. We begin using only the experimentally minimal cuts. Then we find kinematic cuts, such as on rapidity differences, which further purify the samples. Section 2 describes the event samples

and Section 3 the purification procedure. Section 4 discussed theoretical issues associated with defining quark and gluon fractions in perturbation theory. Section 5 summarizes the results.

2. Starting Samples to Explore and Purify

All events were generated with MADGRAPH V4.4.26 [13], a tree-level matrix element generator, using leading order CTEQ6L1 PDFs [14]. Working only at tree-level makes our results independent of any jet-algorithm and showering/hadronization routine. Of course, we do not expect the efficiencies we find to agree with efficiencies one would get after full simulation, or in data, but this is a simple and informative way to determine where quark and gluon jets can be found.

For each sample and each p_T, 200,000 events were generated with the following cuts:

- $p_T^j > p_T$ for all 'jets', meaning any gluons or uds quarks.
- $p_T^\gamma > 20\,\text{GeV}$ for any photons
- $p_T^\ell > 20\,\text{GeV}$ for any leptons from W or Z decays (including missing E_T from neutrinos)
- $p_T^b > 20\,\text{GeV}$ for any b quarks.
- $|\eta| < 2.5$ for any jet, b, photon, or charged lepton.
- $\Delta R > 1.0$ between any two jets.
- $\Delta R > 0.5$ between any jet and any photon or between any jet and charged lepton.

Since the quark and gluon fractions, as well as jet properties, can be strongly dependent on p_T, we have to be careful about how we divide the sample into different p_T bins. We will often find that it is the softest jet in a sample, such as the softest jet in the 3jet or γ+2jet sample, which leads to the highest purity. Since the cross sections fall rapidly with p_T, the majority of events for a given p_T cut will fall around that minimum p_T. This is why *all* jets in a given sample must be above the given p_T, with 'jet' here referring only to light quarks or gluons. In the 200 GeV bjj sample, for example, each light quark or gluon is required to have a $p_T \geq 200\,\text{GeV}$, but the b is only required to have a $p_T \geq 20\,\text{GeV}$. In 2-object final states like γ+jet, the γ automatically also satisfies the jet p_T requirement.

Samples where only one jet satisfies the hard p_T cut, with the others having a $p_T > 20\,\text{GeV}$ cut, were also examined. These have larger cross sections, but only the hardest jet tends to fall within the p_T range of interest, and the kinematic cuts required to achieve high purities reduced the cross section below the softest-jet samples discussed here.

The starting cross sections are shown in Figure 1, as a function of the p_T cut applied to all light quarks and gluons. along with the other cuts listed above. If a sample has a bigger starting cross section, it will be able to suffer harder purification cuts while retaining a

substantial number of events. In this plot, the $t\bar{t}$ sample includes the semi-leptonic branching ratios (2 leptons, 2 b's and 2 light quarks) and has the p_T cut applied to only one of the light quark jets. Despite this looser cut, the cross section drops precipitously above 200 GeV, mostly due to the requirement that the jets be separated by $\Delta R \geq 1$. Since the semi-leptonic $t\bar{t}$ cross section is very small compared to the other processes, we conclude that $t\bar{t}$ events are not a good way to get a large quark jet sample, despite the fact that jets coming from the hadronic W decay are 100% quark.

Instead of putting a cut on the p_T of all the jets, we also tried sorting jets by their rapidity. For example, we asked how often the most (or least) central jet is initiated by a particular parton. This was never more effective at purification than sorting by p_T. Rapidity differences will be used to further purify the samples, but for the starting distributions, we stick with the p_T cut.

In the following, 'quark jet' will always mean only u, d, and s quarks. Any b's and c's are treated as perfectly taggable, although it is straightforward to put in the tagging efficiencies. In Figures 2 through 4, we show the fraction of quarks and gluons produced in the various samples as a function of p_T. When dijet events are referred to as 'QG', that means one jet is a gluon and the other is always a uds quark. The fraction of dijet events that are 'QG' does not include cases with b or c jets in the numerator or the denominator.

In Figure 5, we show the probability that a given jet is a quark or gluon as a function of p_T for the different samples, assuming one jet is picked at random. We see that γ+1jet or W/Z+2jets are good for quark jets, and b+2jets or the 3 or 4-jet samples are good for gluon jets. Again, this is just for the generic cuts listed above, and we have not yet attempted to purify the samples using rapidity or other kinematic information.

In order to purify the samples, we can go two ways. One approach is to reject events so that all of the jets in the remaining events have either all quark jets or all gluon jets. In the top panels of Figure 6, we show the fraction of events where *all* jets are quark or gluon. Note that the vertical axis in these plots is logarithmic. The other approach is to look at particular jets in an event, eventually hoping to apply kinematic cuts to purify the quark or gluon content of *that* jet. (Such cuts are the topic of the next section.) In the bottom of Figure 6, we show the fraction where the hardest or softest jet is quark. These starting points indicate that quark jets will be easier to purify than gluon jets.

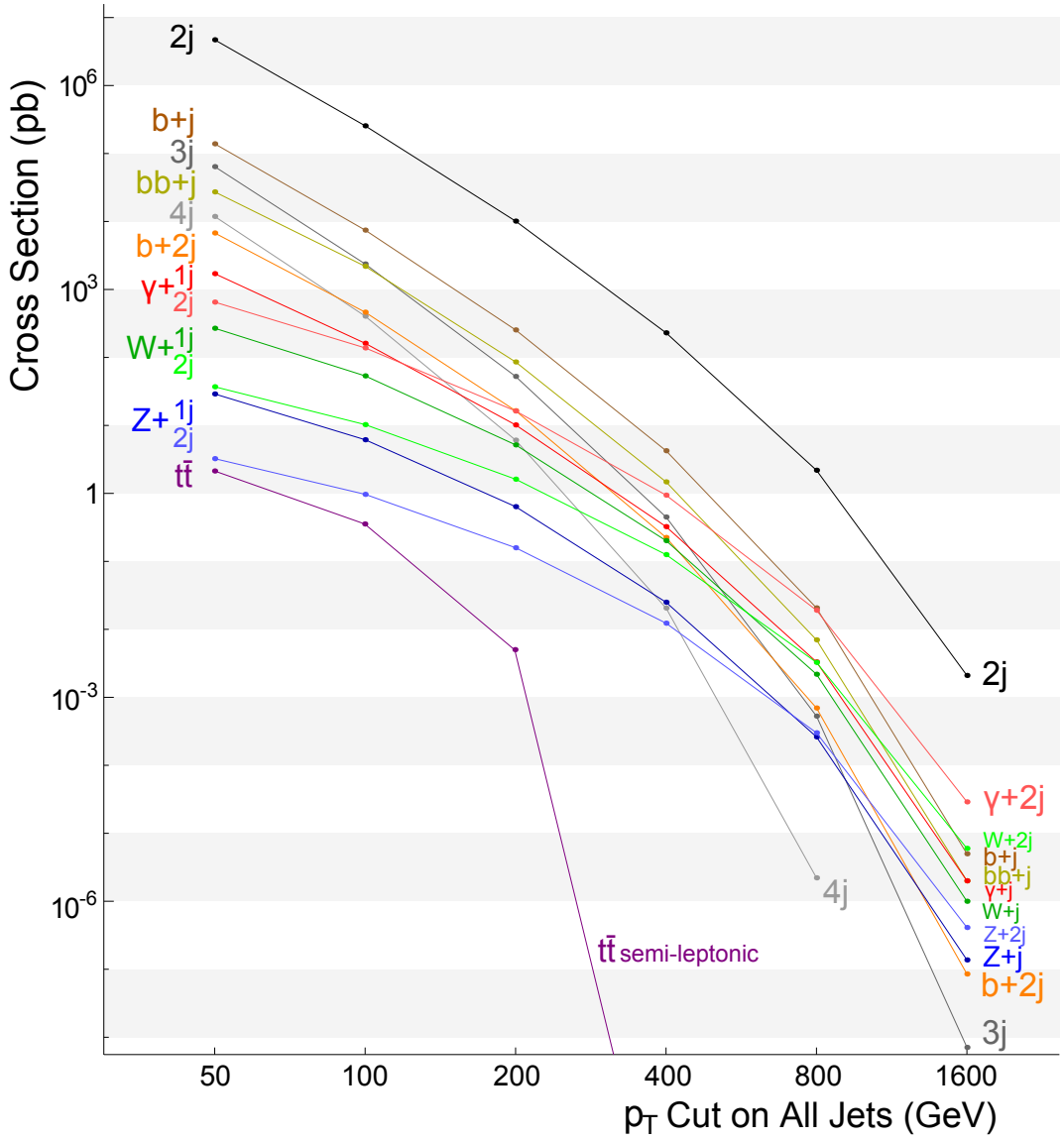

Figure 1: Leading order cross sections, including kinematic cuts and branching ratios for Z/W decay to include an electron or muon. The x-axis indicates the p_T cut applied to *all* light quarks and gluons, but not b-quarks. The constraint on the p_T for b's, photons, and charged leptons or neutrinos from Z/W (though not the Z/W itself) is fixed at 20 GeV. Note that the 3-jet cross section falls below b+2jets due to the harder cuts on the non-b jets. The $t\bar{t}$ cross section refers to the semi-leptonic sample, and, in contrast to all the other samples, the p_T cut is applied to only one of the two light-quark jets. Since its cross section is so low, it will not be considered further.

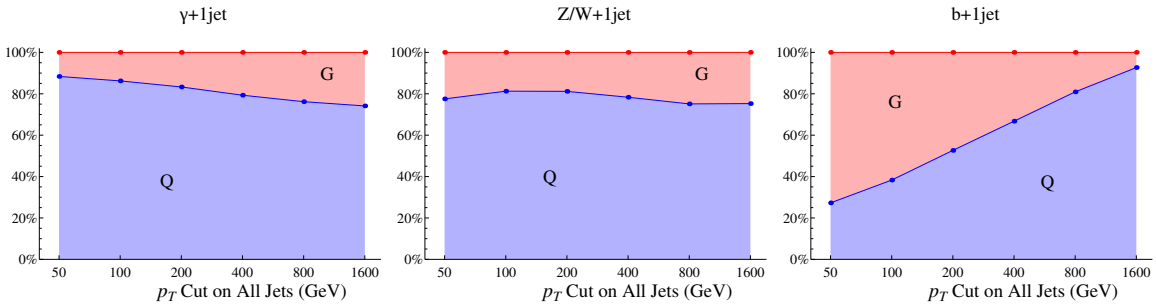

Figure 2: Fraction of X+1jet events where the jet is *uds* quark (bottom and blue in each plot) as compared to gluon (top and red). The horizontal axis is a p_T cut on the jet, which in these events translates into an identical p_T cut on the other object.

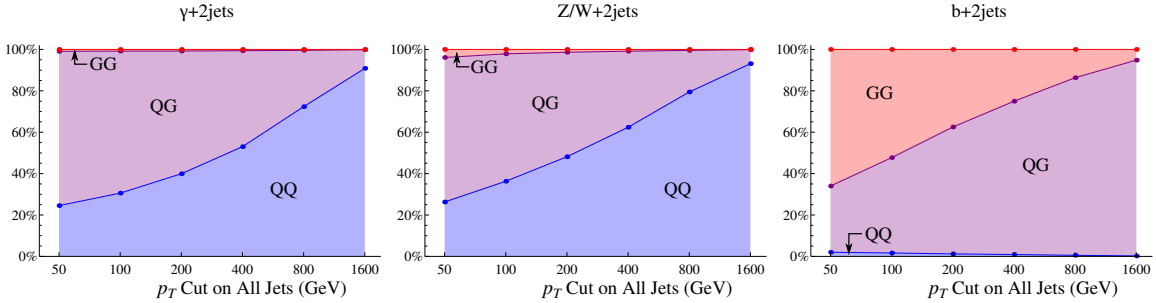

Figure 3: Fraction of X+2jet events where the jets are both light quark 'QQ' (bottom blue) vs one light quark one gluon 'QG' (middle purple) vs both gluon 'GG' (top red). Notice $\gamma + GG$ almost never happens, nor does $b + QQ$. These are starting points for quark and gluon purification. The horizontal axis is a p_T cut on all jets, while the other objects (b, γ, and leptons from Z/W) have $p_T > 20\,\text{GeV}$.

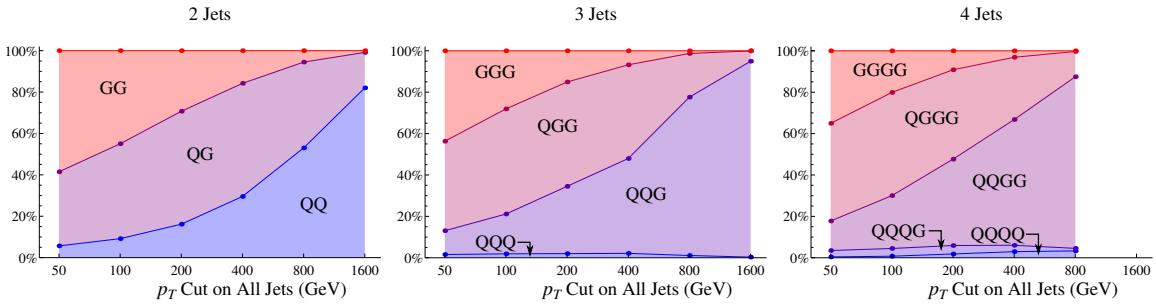

Figure 4: Division of the multijet (dominantly QCD) sample. The horizontal axis is a p_T cut on all jets. Notice that all three jets are almost never all quark, and in the 4-jet sample, there are almost always at least two gluons. The 3-jet sample will be a staring point for gluon purification.

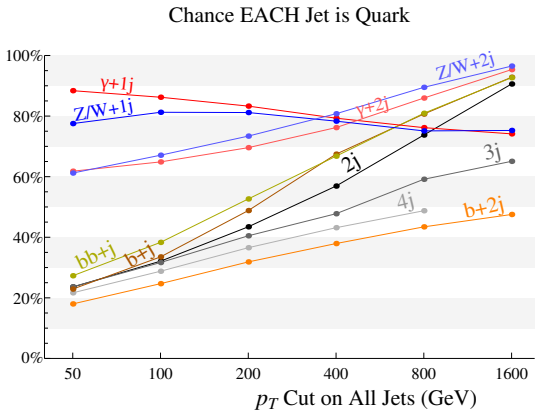

Figure 5: The chance that a given jet is a light quark jet rather than a gluon jet. (This ratio does not include bottom or charm.) The W and Z were nearly identical and combined on this plot, but they are slightly different from the photon, mostly due to the γ and lepton cuts.

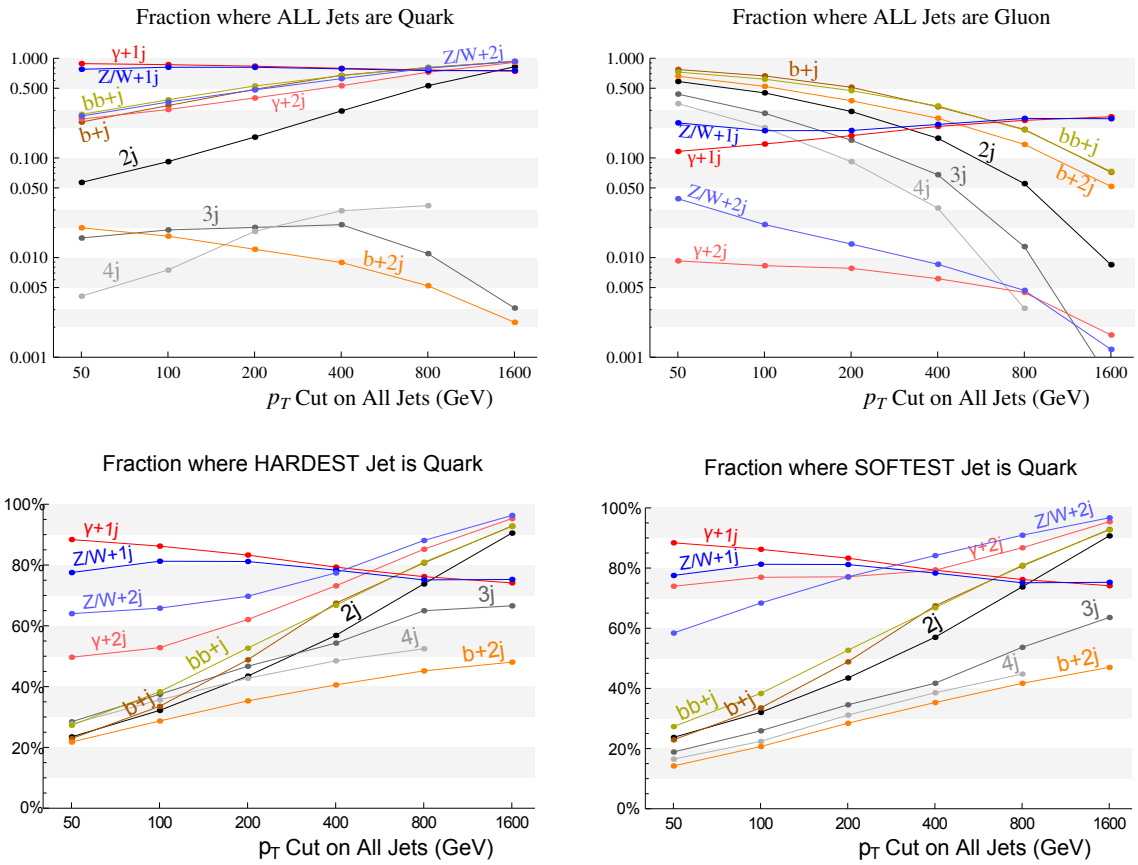

Figure 6: The top row shows the fraction of events where *all* jets are quark or gluon, on a log scale. The bottom row shows the fraction where the *highest* p_T jet is quark, and where the *lowest* p_T jet is quark, on a linear scale. (One minus this fraction are gluon jets.) Having more jets allows for more kinematic handles and potentially better purity.

3. Purifying the samples

In this section, we consider how to improve the purity by judicious kinematic cuts. It's actually quite challenging to get high purities, as we will see. For example, if you start with a 50% pure quark sample and you find a set of cuts that reject two gluons for every quark kept, your new purity is *not* 75%, but only 66%. To reach 75%, you need a cut that rejects three gluons for every quark.

Any cut will have some efficiency ε_q to keep quark jets and a different efficiency ε_g to keep gluon jets. Let q be the starting fraction of events where the jet in question (e.g. the lower p_T 'softer' jet) is a light quark, and $g = 1 - q$ the fraction of events where it is a gluon. Then, after a cut,

$$q \;= \frac{q}{q+g} \;\stackrel{cut}{\to}\; \frac{q\varepsilon_q}{q\varepsilon_q + g\varepsilon_g} \;=\; 1 \Big/ \left(1 + \frac{g}{q}\frac{\varepsilon_g}{\varepsilon_q}\right) \;=\; q_{\text{new}} \tag{3.1}$$

Say we want to optimize the quark purity. One particular cut on the set of kinematic variables will be the best cut for a *particular* quark efficiency ε_q. This will be the cut that lowers the gluon acceptance ε_g as much as possible.

To reach a given quark purity, it obviously helps to start with a sample that's mostly quark. But it is possible to find effective kinematic cuts that improve a mediocre quark purity. This is the case in the γ+2jet sample. Strong cuts can increase the quark purity quite a bit for some samples, but at the cost of a much lower cross section. In the following, we will be careful to express our results as the cross section for quark and gluon jets with a given purity.

3.1 Quark jet purification

We begin by discussing purifying quark jets. As can be seen in Figure 2, the γ+1jet sample appears to be a good starting point, with roughly 80% quarks. This fraction is just the fraction of direct photons produced in the annihilation channel $q\bar{q} \to g\gamma$ (20%) versus the Compton channel $qg \to q\gamma$ (80%), which is in turn set by the gluon and $\bar{q}$ PDFs. Since the gluon PDF is larger than the $\bar{q}$ PDF in a proton, the Compton channel dominates. Unfortunately, the 1-jet samples, such as γ+1jet or W/Z+1jet, do not leave many options for kinematic cuts. Rapidity cuts do not do much, since at high p_T, the jets are more-or-less central, and the cross sections are basically fixed by the PDFs. In fact, the quark purity saturates at roughly 88%. Thus, is helpful to have additional jets to get an additional handle on the kinematics, which will lead us to purities approaching 100%.

We turn next to the the next best sample, γ+2jets. Note that W/Z+2jets is kinematically very similar, but since it has a smaller starting cross section, we focus on the photon. The rapidity distributions for the photon and the softer and harder jets in the samples are shown (for $p_T \gtrsim 200\,\text{GeV}$) in Figure 7. These 1D distributions look like they contain some information, but there is in fact more information in their correlations. Figure 8 shows the 2D distribution of the rapidity of the harder jet and the rapidity of the photon. The likelihood map constructed from these distributions is shown in the third panel. Contours of constant

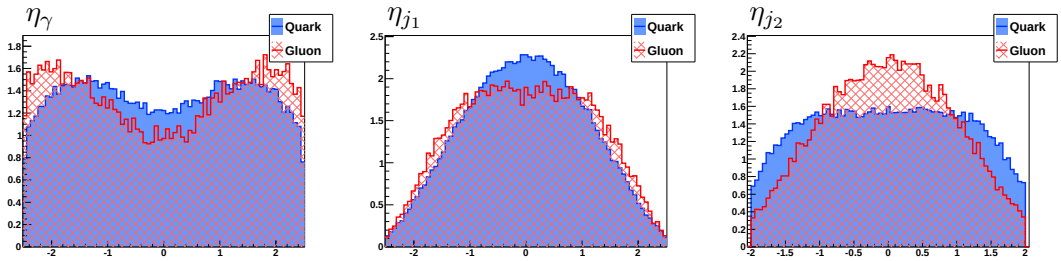

Figure 7: To purify quarks, the best starting point is the softer jet in the γ+2jet sample. The η of the photon (**left**) along with the harder (**center**) and softer (**right**) jets look different when the *softer* jet is a quark (blue solid) vs a gluon (red hashed). These distributions are normalized to equal area. (200 GeV sample shown)

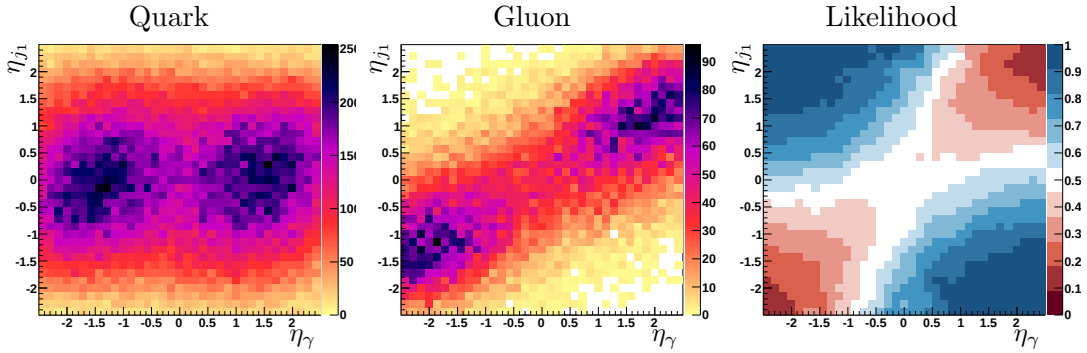

Figure 8: For the quark-heavy γ+2jet sample, a 2D version of last figure's first two histograms: η_γ of the photon vs η_{j_1} of the harder jet. The **left** histogram is for when the softer jet is a quark, and the **center** histogram is for when the softer jet is a gluon. Though we are trying to purify the *softer* jet, it's best to cut on η_γ and η_{j_1} of the *harder* jet. ¿From the left histogram it's clear that when the softer jet is a quark, the harder jet is quite central and the photon's $|\eta|$ is higher and uncorrelated. When the softer jet is a gluon, the harder jet is often toward the edge of our η_j cut, with the photon nearby in η. Correlations are lost of one takes the absolute value of these ηs. The likelihood ratio on the **right** combines each bin as $q/(q+g)$, with blue being more quark-like. When the photon and harder jet are widely separated in η, the softer jet is likely quark. (200 GeV sample shown)

likelihood are very well approximated as contours where the product of the rapidities $\eta_\gamma \eta_{j1}$ is constant, as shown in Figure 9. The quark/gluon discriminant for this product variable is also shown in Figure 9. It clearly has more discrimination power than any of the individual rapidities.

Another option for the γ+2jet sample is to consider the ΔR's between the photon and the jets. Due to a collinear singularity in $q \to q\gamma$, it is natural to expect the photon to be close to one of the quarks. This is in contrast to the gluon case, since there is no $g \to g\gamma$ vertex. The distribution of ΔR between the photon and each jet is shown in Figure 10. Performing a similar 2D likelihood analysis as with just the rapidity inputs, we find the that the single variable $\eta_\gamma \eta_{j1} + \Delta R_{\gamma j_2}$ does very well. Its distribution is also shown.

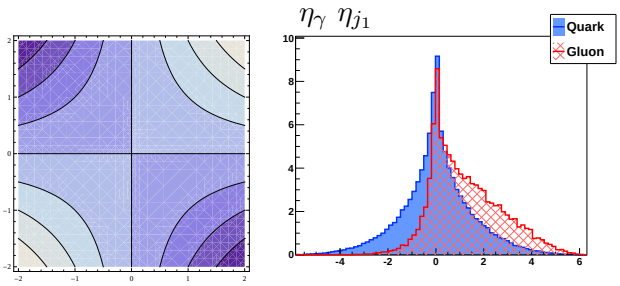

Figure 9: Cutting on any contour in the 2D likelihood distribution above is statistically the optimal discriminant for each quark efficiency, given only these two variables. The contours are roughly given by $\eta_\gamma\ \eta_{j_1}$, which is plotted here on the **left**. We will see that this single variable captures most of the discrimination power of the full 9D multivariate likelihood estimate. On the **right** is the distribution of this product. (200 GeV sample shown)

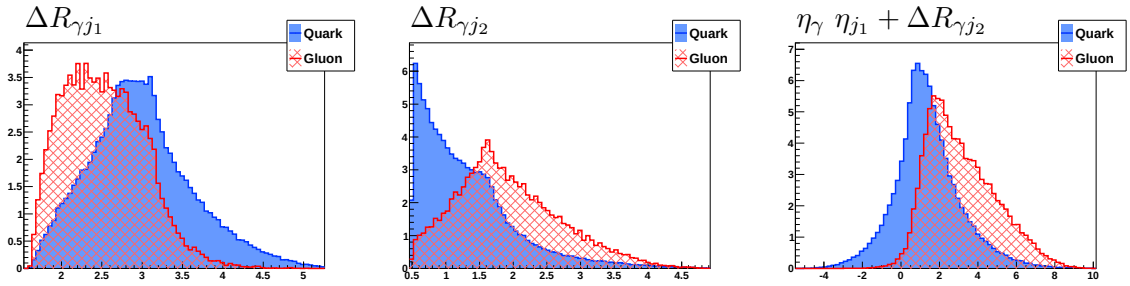

Figure 10: For the quark-heavy γ+2jet sample, distance between the photon and the harder jet (**left**) and softer jet (**center**). Notice that the photon is often as collinear with the softer jet as our $\Delta R_{\gamma j_2} > 0.5$ cut allows. Doing the same 2D likelihood examination as before, an even better single variable discriminant is: $\eta_\gamma\ \eta_{j_1} + \Delta R_{\gamma j_2}$, a combination of the product of η of the photon and *harder* jet, plus the distance to the *softer* jet. The distribution of this mixed variable is shown on the **right**. (200 GeV sample shown)

In constructing unusual variables like $\eta_\gamma \eta_{j1} + \Delta R_{\gamma j_2}$, it is natural to wonder if we are being sufficiently comprehensive. Considering that for a sample with n final-state on-shell quarks and gluons, there are only $3n$ degrees of freedom, it is possible simply to put these 6, 9 or 12 variables into a multivariate analysis. (Transverse momentum conservation and rotational symmetry can reduce the number of degrees of freedom by 3, but it does not hurt to include some redundant information.) More precisely, we input the (p_T, η, ϕ) of each object at a Boosted Decision Tree, which is easy to do with TMVA [15] package for ROOT [16]. The results can be taken as a best case, to which our single variable cuts can be compared. (To be honest, we arrived at this single variable partly by observing which variables TMVA found most important).

The results of the multivariate analysis for quark jet purification are shown in Figure 11. On the left side is the results for 200 GeV jets, cutting on the BDT output. Note that, as

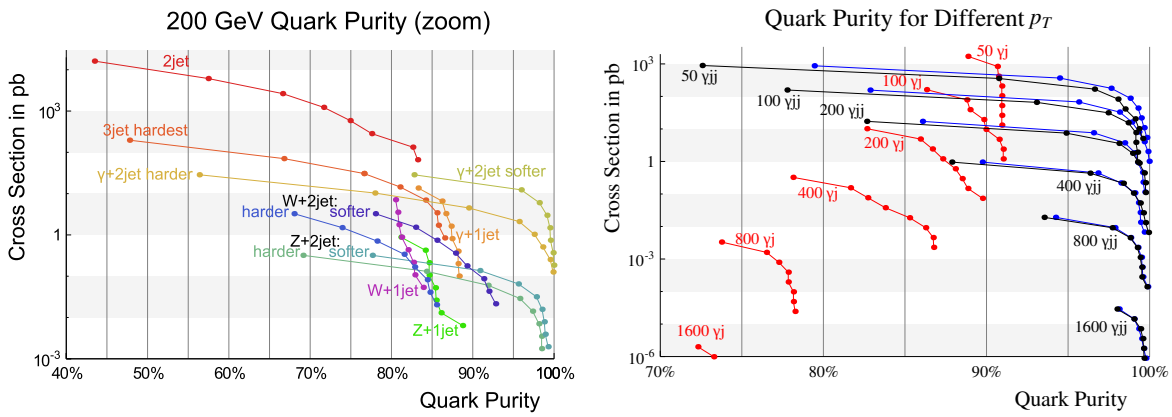

Figure 11: Cross section as a function of quark purity. The **left** panel shows the purity for the different samples with a 200 GeV cut on all non-b jets. The different points correspond to different cuts placed on a Boosted Decision Tree output, trained to optimize the quark purity. The leftmost dots of each sample are the uncut purities, and each successive dot corresponds to cutting the number of events in half. By the final dot, which keeps $1/128^{\text{th}}$ of the signal, cutting harder no longer increases the purity. The **right** panel shows the purities for the γ+1jet (red) and γ+2jet (blue) samples for various p_T's, where the cuts are with BDTs trained on 6 and 9 kinematic variables, respectively. The black curves correspond to purities obtained after cutting on the single variable $\eta_\gamma\eta_{j_1} + \Delta R_{\gamma j_2}$. The blue curve takes the jet closest to the photon as a starting point, whereas the black curve takes the softer of the two jets as its starting point. This is the reason for the lower initial purity but the same cross section. (It was easier to find a single variable using the softer jet rather than the jet closer to the photon.)

anticipated, the γ+1jet cannot be purified much — putting harsher cuts hits a wall and eventually just kills the cross section. On the right, we focus on just the γ+1jet and γ+2jet samples for all p_T. The red curves are the BDT output using 6 inputs for γ+1jet, the blue curves BDT with 9 inputs for γ+2jets, and the black curves for our single variable $\eta_\gamma\eta_{j1} + \Delta R_{\gamma j_2}$. It is nice that the single variable does as well as the comprehensive analysis using the 9 BDT inputs.

We conclude that the best way to get a clean quark sample at low p_T is to use γ+1jet, for simplicity, or γ+2jets at moderate to large p_T, cutting on the single variable $\eta_\gamma\eta_{j1} + \Delta R_{\gamma j_2}$. Depending on how much cross section you are willing to sacrifice, for 200 GeV jets, you can get 95% quark purity at 2 pb or 99% purity at 500 nb.

3.2 Gluon jet purification

Next, we turn to the more difficult case of gluon jet purification. It is more difficult because there is no starting sample with purity above 80%, and because there are no simple physically motivated handles for purification. Indeed, for the quark, we used the fact that there is a collinear $q\gamma$ singularity but no $g\gamma$ singularity to inspire a $\Delta R_{j\gamma}$ cut. But for a gluon we cannot use the gq singularity since we are trying to avoid q jets all together. The exception is

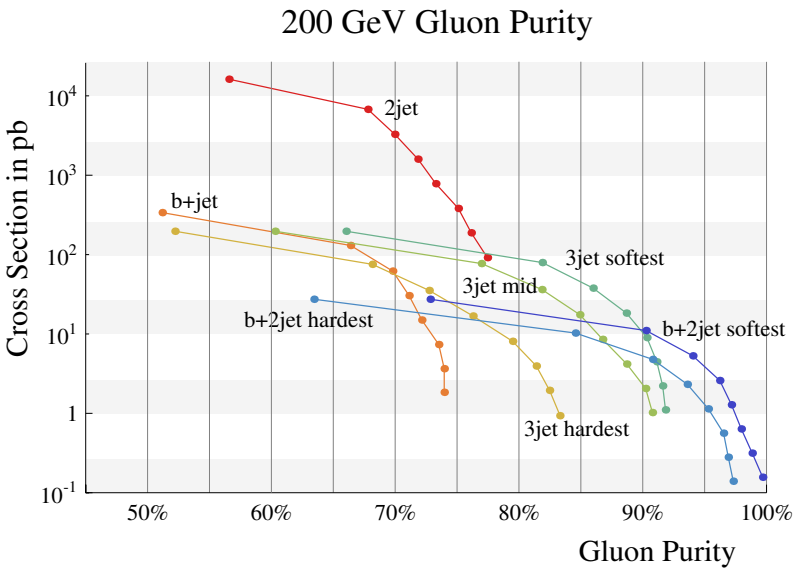

Figure 12: Cross section as a function of gluon purity for the different samples with a 200 GeV cut on all non-b jets. The different points correspond to different cuts placed on a Boosted Decision Tree trained to optimize the gluon purity. The leftmost dots of each sample are the uncut purities. There are 3 curves for the 3-jet samples, and two for the b+2jet samples, corresponding to which of the jets (from hardest to softest) is being considered. Note the three 3-jet samples start with identical cross sections, but higher purities are achievable for the softer jets.

samples with jets and b's, where we can use b-tagging information to help purify the sample. This will in fact be relevant, but we will find that the 3- and 4-jet samples actually work quite well, and avoid having to deal with b-tagging.

To begin, we start with a multivariate BDT analysis using as inputs the (p_T, η, ϕ) of all final state particles. The results for the different 200 GeV samples are shown in Figure 12. We can see that while the b+2jets has good efficiency, it also has a cross section orders of magnitude smaller than the 2-jet sample. The 3-jet sample is somewhere in between, with efficiencies about 80% for a cross section of 100 pb. We will consider these three samples in the following, as there may be situations when each one is advantageous.

First, consider the b+2jet sample. Looking back at Figure 3, we see that there is a contribution from both 'GG' (with ggb final states) and 'QG' (with qgb final states). The ggb section obviously has perfect gluon efficiency regardless of cuts. The main parton level process contribution in the qgb channel is $ub \to ubg$, which looks like final state gluon radiation from t-channel $ub \to ub$. Since we put a harder cut on the u and g than the b, the kinematics will mostly have the u going back-to-back with the gb, and so the g will be somewhat softer. This explains why the starting efficiencies for the softer jet at p_T=200 GeV are around 73%, versus 63% for the harder jet, as shown in see Figure 12.

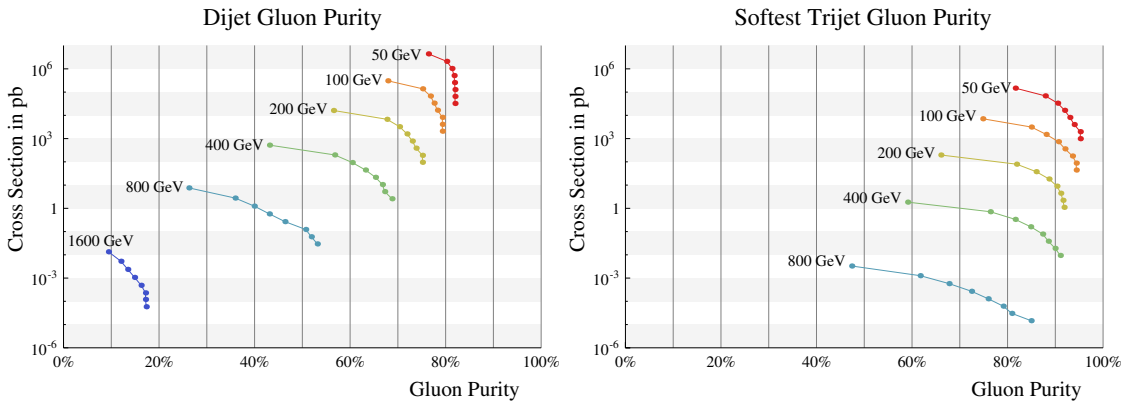

Figure 13: Gluon purities for the dijet and trijet samples for different p_T's. For each p_T sample, the first dot on the top-left represents the starting purity and cross section with no kinematic cuts.

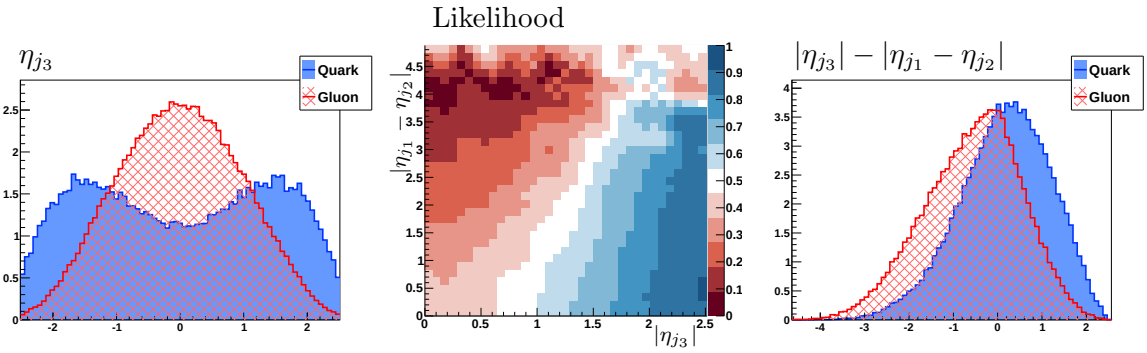

Figure 14: To purify gluons in the 3-jet sample, we look at the softest jet, which tends to be central. It's η is shown on the **left**. An even better discriminant takes into account the separation of the harder two jets and the correlation between this separation and the softest jet's η is shown in the **center**. A good *single* variable capturing the likelihood contours is $|\eta_{j3}| - |\eta_{j1} - \eta_{j2}|$ whose distribution is shown on the **right**. (200 GeV sample shown)

The main complication in the b+jets samples is efficient b-tagging. So far, we have assumed perfect b-tagging, so that both jets are effectively anti-b-tagged. In reality, b-tagging can be made very tight, keeping only jets that really look like b-jets or really look like non-b-jets. A very tight b-tag will lower the cross section without affecting the purities shown. If looser b-tagging is used, the cross section will be higher but mistags of jjj and mis-anti-tags of bbj make the analysis more complicated. Note, however, that the dominant background to b-jets are charm jets and from the point of view of finding gluon jets, it is ok to treat charm jets as b-jets. In many ways b-jets act like gluon jets rather than like light quark jets. For example, the OPAL experiment at LEP [17] found b-jets to have more charged particles over a wider area than light quark jets, making them similar to gluon jets in this regard. It is therefore very important to have tight anti-b-tagging on any jet used in further

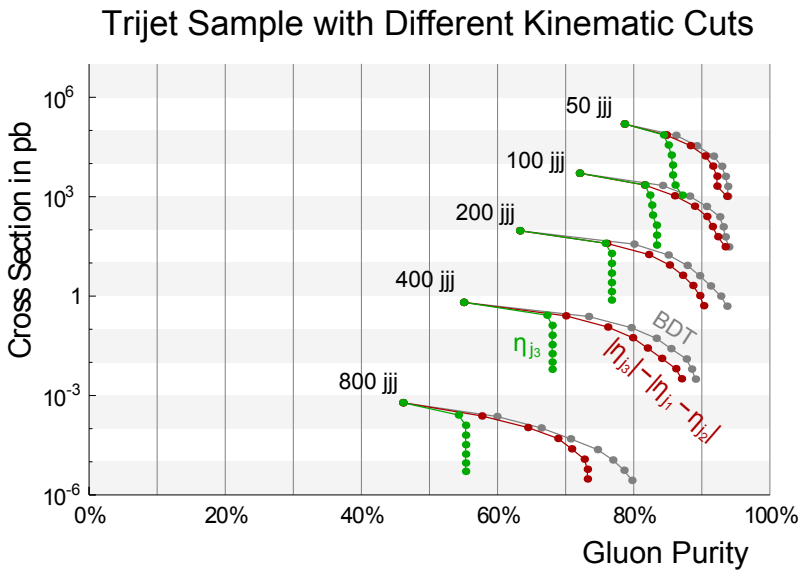

Figure 15: Cross section as a function of gluon purity. The gray curve shows how pure the sample could be made using all the kinematic information, through a Boosted Decision Tree. The green curves show the result of cutting on the rapidity of the softest jet η_{j_3}. The red curve shows that by cutting on the single variable $|\eta_{j_3}| - |\eta_{j_1} - \eta_{j_2}|$, nearly optimal purities can be achieved, matching the BDT. Note, all three curves agree at their left-most points, where no cut is applied.

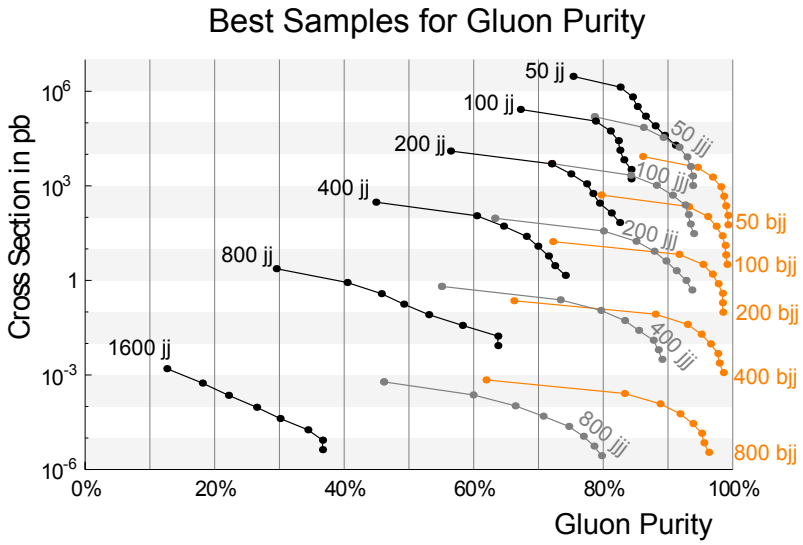

Figure 16: Cross section as a function of gluon purity. The gluon tagging efficiencies for the dijets (black), trijets (gray), and b+2jets (orange) are shown. All curves correspond to the result of an optimal purification using a multivariate analysis. Nearly optimal results can be reproduced in the trijet variable with a simple cut on a single kinematical variable, as described above.

analysis, no matter which starting sample it came from. Since b-tagging is very detector and p_T dependent, we do not attempt to include it in any quantitative way in this tree-level study.

Next, consider the dijet and trijet samples. There is actually a fairly strong p_T dependence in the gluon fractions, as can be seen in Figure 4. As before, we begin by using full kinematic information in Boosted Decision Trees. The result is shown in Figure 13. We see that dijets have a higher cross section, but cannot be purified beyond a limiting value. The trijet sample can be purified more, but has a lower cross section since its softest jet must be above the indicated p_T. While the efficiencies are not as high as in b+2jets, the trijet sample can provide 90% gluon purity with large cross sections and few b-tagging worries. A similar analysis can simplify the kinematic cuts to a few variables.

The best single simple variable to cut on for the softest jet in the trijet sample is the rapidity of that jet, η_{j3}. Its distribution is shown in the left panel of of Figure 14, where we can see that the softest jet tends to be central when it is a gluon and more forward when it is a quark. Unfortunately just cutting on the rapidity of the softest jet can only do so well in purifying the sample. This can be seen from the distributions – there is no region which is pure gluon. To be more quantitative, the effect of cutting on η_{j3} is shown in Figure 15. The green, representing cuts on η_{j3} hits a hard wall for each p_T.

To progress further, we observe that η_{j3} is only weakly correlated with the rapidity difference of the other two jets, $|\eta_{j2} - \eta_{j1}|$. The 2D distribution and the likelihood contours are shown in the center of Figure 14. These contours are well mapped by $|\eta_{j3}| - |\eta_{j2} - \eta_{j1}|$, which we take as our best composite variable. Its distribution is shown on the right of this figure. Note that, in contrast to η_{j3}, the distribution of this composite variable has a gluon tail toward negative values. Thus, it should be possible to put very hard cuts on it to improve efficiency. The result is shown and contrasted to the full BDT and η_{j3} results in Figure 15. We see that cutting on this variable does nearly as well as using the full kinematic information.

The results for the dijet, trijet and b+2jet samples are summarized in Figure 16. To get very high $\sim$99% gluon efficiencies, one needs the b+2jet samples with excellent b-tagging. But at 80% or 90%, one can instead use trijets cutting on the discriminant $|\eta_{j3}| - |\eta_{j2} - \eta_{j1}|$. The trijet sample has a much larger cross section than b+2jets for the lower jet p_T samples.

4. Defining quark and gluon jets in QCD

In this section, we discuss what exactly is meant by quark and gluon jets. We begin by considering particle decays, since they provide a context in which the concept of quark and gluon jets is more intuitive. We then discuss how soft and collinear radiation preserves the identity of a jet as quark or gluon, and how quark and gluon cross sections can be defined beyond leading order.

Consider a Z boson which decays to 2 jets. In the limit that the jets are highly collimated and well separated, these jets are 100% quark jets. This is not to say that there are no gluons represented in the jets — beyond leading order in perturbation theory there will be many gluons, and these gluons can have as much energy, or more, than the quarks — but the *jets*

coming from the Z-decay are still quark jets, by definition. (There is actually zero probability for the jets to be gluon jets in this case due to Yang's theorem.) One could also imagine a particle which would decay only to gluon jets, for example, a light Higgs boson that only couples directly to the top (the decay would be through a top-loop). Here, the jets would unambiguously be 100% gluon jets. If a particle decays to 3 jets, one can ask about the quark and gluon content of the third jet as well. This would also be well-defined to the extent that the jets are collimated and separated, which is the same extent that the jets are representative of the hard interaction at all. In a multiparticle cascade decay with many jets, such as in supersymmetry, one can also ask unambiguously about the quark or gluon jet content of the various jets produced. In fact, even in QCD processes, such as $pp \to$dijets the concept of quark and gluon jets is no more ambiguous than in decays, one is just less used to thinking about quark and gluon fractions.

When jets are highly collimated and well separated, their cross sections factorize into the production process, for which there is no mixing between quarks and gluons, and the fragmentation process, whereby those quark and gluon jets shower and hadronize into observable particles. Although exact factorization proofs are not available for anything but the simplest process (Drell-Yan), scaling arguments suggest that any violations to factorization should be negligible. Thus, the concept of quark and gluon jets is a well-defined theoretical concept up to power corrections that scale as $\Lambda_{\rm QCD}/E$ and $R \sim m/E$, where R is the size of the jet, E its energy and m its mass.

As mentioned in the introduction, there is no ambiguity at leading order in defining the fraction of quark and gluon jets in any exclusive sample. To be precise, leading order here means the Born level, the lowest order in perturbation theory which produces the required number of jets. To be concrete, consider for example the direct production of a hard photon, say with $p_T > 200$ GeV. At leading order, there are two Feynman diagrams, the Compton channel: $qg \to q\gamma$ and the annihilation channel $q\bar{q} \to g\gamma$. The ratio of the cross sections for these channels, at leading order, tells us that 85% of the jets produced in association with a photon will be quark jets. For more complicated processes there is also no ambiguity as long as we are specific about which jet we mean, in an infrared safe way. For example, we can ask about the 2nd hardest anti-k_T $R = 0.4$ jet in W+jets events. Here, the Born level is W+2 jets, and the cross section ratio can be computed unambiguously (up to scale uncertainties) at leading order.

At next-to-leading order, there are virtual and real contributions. Both of these are infrared divergent and some part of the real contributions must be added to the virtual to get a finite answer. The virtual graphs have the same number of jets as the Born level, and so whether they contribute to the quark or gluon jet cross section is similarly unambiguous. The real graphs can be split into a contribution containing the infrared divergent regions and a hard remainder. The infrared divergences are soft or collinear, and in either limit the identity of the jet as quark or gluon is conserved. In the soft limit, the interactions of gluons are Eikonal and factorize off, again leaving the quark or gluon nature of the jet unchanged. In the collinear limit, helicity is conserved. So one can treat the helicity of a jet as a conserved

quantum number which is necessarily different for quark and gluon jets. Moreover, for any infrared-safe jet definition, a collinear gluon emitted in the singular region must go into the jet, so the overall baryon number of the jet (number of quarks minus number of antiquarks) is conserved. Hard emissions must produce another jet, at least in the approximation where the jets are highly collimated, which is where factorization holds.[1] So the infrared-singular parts of the real emission contributions do not change whether the jet is quark or gluon and therefore the quark or gluon fraction can be defined at higher orders in perturbation theory.

To all orders in perturbation theory, the factorization into quark and gluon production can be simplified by the use of operators in Soft-Collinear Effective Theory [21, 22]. For example, for direct photon production [23], there are 6 production channels, with initial states $qq, \bar{q}\bar{q}, q\bar{q}, qg, gg$ and $\bar{q}g$. Each channel has two spin structures, corresponding to the cases when the quarks have equal or opposite spin. For example, in the $q\bar{q} \to g\gamma$ channel, the operators are

$$\mathcal{O}^{S\nu}_{q\bar{q}} = \bar{\chi}_2 \mathcal{A}^{\nu}_{\perp} \chi_1, \qquad \mathcal{O}^{T\nu}_{q\bar{q}} = \bar{\chi}_2 \sigma_{\mu\nu} \mathcal{A}^{\mu}_{\perp} \chi_1, \tag{4.1}$$

So there are 12 operators total relevant for matching at the Born level. The fields χ and $\mathcal{A}$ are collinear quarks and gluons with associated collinear Wilson lines. For simplicity, these are called jet fields. More details can of the notation can be found in [23].

The point of the SCET notation is that it gives a precise definition to what we have been calling quark and gluon jets. It therefore lets us define the quark and gluon jet fractions exactly, as ratios of matrix elements of operators with quark or gluon jet fields. In the limit where factorization holds, there is no mixing between operators with different jet fields, or even of fields with different spin. For example, in direct photon, when the photon is very energetic there is only phase space for it to recoil against a single jet. In this limit, the process is exactly described by the operators in Eq. (4.1) and the other 10 operators for the other channels. The mixing between the operators is power suppressed. To add some concreteness to the discussion, at leading order, the jet recoiling against at 300 GeV photon is 82.3% quark. At NLO, it is 84.6% quark and at NNLO 85.1% quark. The leading order prediction is a very good approximation to more precise values, since the radiative corrections largely drop out of the fraction.

In summary, in this section we have explained how the quark and gluon jet fraction is exactly defined in a limit in which the production of jets factorizes into an incoherent sum of different channels. This gives precisely calculable cross sections, and hence a well-defined quark-to-gluon jet fraction.

[1] There may be additional "non-global" contributions, from configurations where a hard gluon splits into two quarks and one of those ends up a jet. Whether non-global logs are relevant or not is a question about the observable, such as the jet mass, not about whether the jets are quark or gluon. Quark or gluon jets are defined to the extent that factorization holds, and non-global logs would violate factorization. More information on non-global logs can be found in [18, 19, 20].

5. Conclusions

In this paper, we have systematically explored which processes at a proton collider can be exploited to give pure samples of quark and gluon jets. We found that a 98% pure quark jet sample is achievable by starting with the softer jet in γ+2jets and cutting on the combined kinematic variable $\eta_\gamma \eta_{j_1} + \Delta R_{\gamma j_2}$. The corresponding cross sections are around 10 pb for $p_T \geq 100\,\text{GeV}$, 1 pb for $p_T \geq 200\,\text{GeV}$, or 0.1 pb for $p_T \geq 400\,\text{GeV}$ quark jets. More quark purity information is in Figure 11.

Gluon jets are more difficult to purify. We found that the b+2jets sample provides the best results under ideal conditions. Unfortunately, to get such pure gluon jet samples requires a excellent b-tagger, and a realistic analysis can only be done with details of the particular experiment and b-tagging method. The next best thing, is to use the softest jet in 3jet events. This has a higher cross section than the b+2jets sample, but cannot achieve quite as high purities. Cutting on the combined variable $|\eta_{j3}| - |\eta_{j2} - \eta_{j1}|$, the trijet sample can provide 100 pb at 93% purity for 100 GeV gluon jets, 1 pb for 90% purity 200 GeV jets, or 10 fb of 85% purity 400 GeV jets. More gluon purity information is in Figure 16.

The fraction of quark and gluon jets, which we have calculated in this paper at leading order in perturbation theory, is well-defined theoretical concept, up to power corrections in the jet size. These power corrections are suppressed when the jets are hard and well-separated. The quark-to-gluon jet fraction is a theoretical concept, not directly observable, but it is an extremely useful theoretical concept. The observables are the jet properties in a given sample, which correlate with the quark or gluon jet fraction. These properties, such as mass of the hardest jet, can in principle also be calculated. Certain regions of phase space, the ones with pure samples of quark or gluon jets discussed in this paper, should allow us to test calculations and calibrate simulations of jet properties more efficiently. With the better experimental handle on jet properties arising from the study of these samples, we will be better prepared to extract properties of fundamental standard-model or beyond-the-standard-model physics encoded in hadronic events.

Acknowledgments

We thank Salvatore Rappoccio and Michael Kagan for discussions about CMS and ATLAS, and David Krohn and Peter Wittich for comments on the manuscript. JG thanks Michigan State for its speaking invitation and ensuing discussions that inspired this work, along with a night stuck in the Detroit airport generating samples. This work was supported in part by the Department of Energy under grant DE-SC003916. Computations for this paper were performed on the Odyssey cluster supported by the FAS Research Computing Group at Harvard University.

References

[1] D. E. Kaplan, K. Rehermann, M. D. Schwartz, B. Tweedie, Phys. Rev. Lett. **101**, 142001 (2008). [arXiv:0806.0848 [hep-ph]].

[2] T. Han, D. Krohn, L. -T. Wang, W. Zhu, JHEP **1003**, 082 (2010). [arXiv:0911.3656 [hep-ph]].

[3] J. Thaler, K. Van Tilburg, JHEP **1103**, 015 (2011). [arXiv:1011.2268 [hep-ph]].

[4] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black, B. Tweedie, [arXiv:1010.3698 [hep-ph]].

[5] Y. Cui, Z. Han, M. D. Schwartz, [arXiv:1012.2077 [hep-ph]].

[6] A. Abdesselam, E. B. Kuutmann, U. Bitenc, G. Brooijmans, J. Butterworth, P. Bruckman de Renstrom, D. Buarque Franzosi, R. Buckingham *et al.*, [arXiv:1012.5412 [hep-ph]].

[7] J. M. Butterworth, A. R. Davison, M. Rubin, G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008). [arXiv:0802.2470 [hep-ph]].

[8] D. Krohn, J. Shelton, L. -T. Wang, JHEP **1007**, 041 (2010). [arXiv:0909.3855 [hep-ph]].

[9] A. Banfi, G. P. Salam and G. Zanderighi, Eur. Phys. J. C **47**, 113 (2006) [arXiv:hep-ph/0601139].

[10] J. Gallicchio, M. D. Schwartz, Phys. Rev. Lett. **105**, 022001 (2010). [arXiv:1001.5027 [hep-ph]].

[11] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, C. Lee, JHEP **1011**, 101 (2010). [arXiv:1001.0014 [hep-ph]].

[12] R. Kelley, M. D. Schwartz, H. X. Zhu, [arXiv:1102.0561 [hep-ph]].

[13] J. Alwall *et al.*, "MadGraph/MadEvent v4: The New Web Generation," JHEP **0709**, 028 (2007) [arXiv:0706.2334 [hep-ph]].

[14] D. Stump, J. Huston, J. Pumplin *et al.*, JHEP **0310**, 046 (2003). [hep-ph/0303013].

[15] A. Hoecker *et al*, TMVA Toolkit for Multivariate Data Analysis with ROOT, `http://tmva.sourceforge.net/`.

[16] R. Brun and F. Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also `http://root.cern.ch/`.

[17] O. Biebel [OPAL Collaboration], "A comparison of b and uds quarks to gluon jets," SPIRES Conference C96/08/11.1 (1996)

[18] M. Dasgupta, G. P. Salam, Phys. Lett. **B512**, 323-330 (2001). [hep-ph/0104277].

[19] R. Kelley, R. M. Schabinger, M. D. Schwartz, H. X. Zhu, [arXiv:1105.3676 [hep-ph]].

[20] A. Hornig, C. Lee, I. W. Stewart, J. R. Walsh, S. Zuberi, JHEP **1108**, 054 (2011). [arXiv:1105.4628 [hep-ph]].

[21] C. W. Bauer, S. Fleming, D. Pirjol and I. W. Stewart, Phys. Rev. D **63**, 114020 (2001).

[22] C. W. Bauer, D. Pirjol and I. W. Stewart, Phys. Rev. D **65**, 054022 (2002)

[23] T. Becher and M. D. Schwartz, JHEP **1002**, 040 (2010) [arXiv:0911.0681 [hep-ph]].

[24] J. Gallicchio and M. D. Schwartz, arXiv:1106.3076 [hep-ph].